Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Predicting Attention Drop in Learners Using Interaction-Based Machine Learning Models

Shreetirth Talpallikar[1], Vishnu Bhattiprolu[2], GandlaSrujan[3],Uday Punna[4], Palagati Anusha[5], Ellapu Venkatesh[6]

[1,2,3,4,5]Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India
[6]Assistant Professor, Department of Business Administration, Krishna Chaitanya Institute of Science and Technology, Andhra Pradesh, India.

[1] talpallikar.shreetirth@gmail.com , [2] vishnu.bhattiprolu@gmail.com, [3] srujan8691@gmail.com , [4] udaypunna.in@gmail.com, [5] palagatianushareddy@gmail.com, [6]ellapu.venkatesh9@gmail.com
Corresponding author: ellapu.venkatesh9@gmail.com

Abstract: Digital learning platforms have brought a revolution in the educational world by providing flexible and personalized learning experiences to everyone. The main challenge of these platforms is retaining the learner's attention. The change in attention of the learner results in reduced knowledge gain and learning outcomes. Identifying the core reasons of reduced attention can help learning platforms improve themselves. The purpose of this research is to predict the main reasons and features of the platform for low attention among learners based on the behavioural interaction data obtained from a large database of learners.

Keywords: Attention Prediction, Learning Analytics, Machine Learning, Student Engagement.

Submitted on: 22 April 2026   Accepted on: 05 June 2026   Published on: 10 June 2026

1. INTRODUCTION

This research aims to use the ML models on the basis of interaction data to predict the attention drop of learners in the digital learning environment. The interaction data collected from the interaction process of the learners reveals that behavioral data can be used effectively to predict the decrease of learners' attention. This research does not want to take the results of the learners' feedback but directly focus on the physiological aspects of the learners, but instead proposes the use of simple interaction features which can be easily acquired during the learning process.

Various classification models were used to examine the interaction data collected from the learners to predict the drop in attention among them. Among all the models analyzed in this research the ensemble models were found to have the highest accuracy than the baseline models in understanding the factors associated with attention drop of users. Feature importance analysis was also used to identify important behavioural features that can make the learners have a drop in attention.Idle time and interaction frequency are among these features. This assists in developing learning systems that can adjust the content and make different changes in their platforms to re-engage the learners. Finally, this study shows

DoI: doi.org/10.63949/crinfo.v2i2.001

that ML techniques that utilize interactions are cost-effective and effective ways to augment student engagement and learning in online education.

2. Problem Statement

In digital learning environments, the problem of maintaining learners' attention is a significant challenge that impacts knowledge gain, engagement, and overall learning goals. Unlike the physical learning environments, digital learning platforms are not continuously monitored, and it is difficult to detect when a learner begins to lose attention. In physical learning we generally have a human interaction with the learners and these human interaction makes the learners feel encouraged and enhances engagement.

On the other hand, modern digital learning platforms are generating high-quality interaction data, including session rate, time spent on learning activities, content completion patterns, and test performance. Nevertheless, the application of this data for attention drop detection is still a challenge due to the subtle and volatile nature of human attention.

The first problem that this study aims to solve is the development of an accurate and scalable solution that can predict attention drop using only non-intrusive, interaction-based behavioural features. This problem involves determining whether machine learning algorithms can accurately separate normal and low-attention learners based on their engagement behaviour, even when such data lacks separability and variability.

3. Architecture

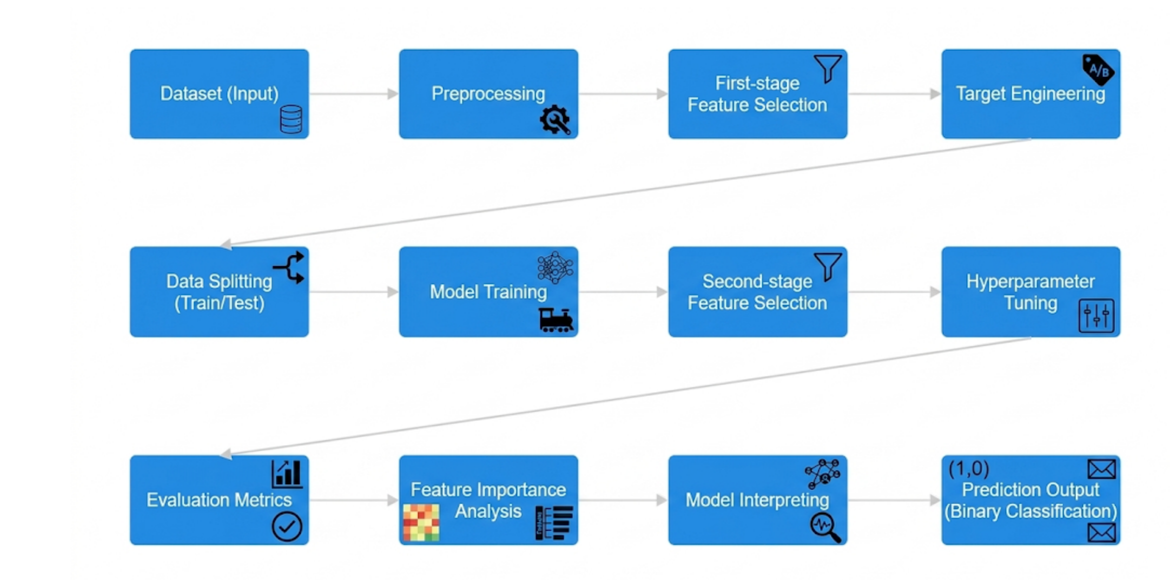

Figure 1. Architecture

The proposed system is a structured machine learning training platform to predict the attention drop-out of learners and based on behavioural data of interaction. The architecture starts with the input of a big data digital learning data comprising of user engagement and performance related features. This raw data undergoes a preprocessing phase which involves missing values and ready the dataset to be processed.

DoI: doi.org/10.63949/crinfo.v2i2.001

Frontiers in Engineering and Informatics
ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

Features related to learners' interest and learning behavior, as well as their academic achievements, are identified after the pre-processing. The engagement consistency feature is then used to create a target variable, based on a percentile based threshold, which is used to create a binary classification problem.

Processed dataset will be split into two part namely training and testing data using stratified sampling to retain the ratio of classes. Various ML models are then built and trained on the training data to enable linear as well as nonlinear learning of patterns. Such common performance measures are part of the evaluation of the models after training. Besides, the analysis of feature importance which is used to determine the major behavioral factors influencing the drop of attention is also done. Lastly, the system generates such forecasts that categorize the learners as low attention or normal attention, which could be utilized by online education programs to develop specific interventions and enhance learning activities.

4. Methodology

4.1 Dataset Description

This study relies on a large digital learning analytics (DLA) dataset containing approximately 1,00,000 learners' records. The dataset includes various behavioural and performance related features that have been collected from an online learning environment from 2020 to 2025. Each of these features has been a measure of users' engagement with the platform, including time spent on the site, frequency of engagement, content completion behavior, and learning outcomes.

4.2 Formulation of the Problem

This study's intention is to use machine learning techniques to predict when learners' attention span declines. Attention is derived from user engagement behaviour because it cannot be directly be observed or evaluated

From the engagement_consistency feature, a target variable, low_attention, is derived to formulate a binary classification problem. The 30th percentile of engagement consistency is used to define a threshold. Low attention (1) is assigned to the learners whose values fall below this cutoff, while normal attention (0) is assigned to the remaining learners. So, (1) represents low attention and (0) represents normal attention

4.3 Selection of Features

Ten features were chosen to depict various aspects of learner behavior and engagement in the digital learning environment. Engagement-related features consisted of the number of sessions per week and the amount of time that was spent using the app daily, both of which represented the level of involvement with the material. The video completion percentage was used to assess content interaction, and learner participation and collaborative engagement were gauged through forum postings. Academic consistency was represented as the proportion of assignments submitted, while total hours learning showed overall learning effort. To measure willingness to reengage with difficult concepts and assess the corrective

DoI: doi.org/10.63949/crinfo.v2i2.001

learning behavior of learners, the number of remediation modules completed was included. Learning efficiency score and mastery score, which are linked to learning effectiveness and knowledge acquisition, were used to characterize performance properties. Plus, content_difficulty_avg is added, which takes into account the difficulty of the content that students come across during the learning process. In sum, these features reflect aspects of behavior, cognition, engagement, and performance, which are of significant interest for modeling and prediction of declines in attention in online learning settings.

4.4 Preprocessing Data

The data were preprocessed before the process of developing the model to improve the quality of data and then increase accuracy in detecting fake reviews. Missing values were imputed with the median, a very strong method that lowers the influence of outliers and maintains the data distribution's overall shape. If required, the feature distribution was normalized by means of feature scaling on all variables that differed in their value ranges. The dataset was preprocessed and split into a training:testing set in the ratio 80:20 to train and evaluate our model. In addition, stratified sampling was used to ensure a similar distribution of classes in each of the data splits, which helps reduce the risk of bias and provides a valid evaluation of the performance of the model.

4.5 Model Creation

This study employed three machine learning algorithms: logistic regression, random forest, and XGBoost, each presenting distinct strengths and weaknesses. Due to its simplicity, efficiency, and interpretability, logistic regression served as the baseline model to explore the linear relationship between learner behavior and attention levels. An ensemble learning approach known as random forest, which constructs multiple decision trees, modeled complex feature interactions and captured non-linear patterns in the data. The model was powered with XGBoost, a highly efficient and useful gradient-boosting method for structured data. A machine learning pipeline consisting of data preprocessing and classification stages was used for each algorithm to ensure a consistent and reproducible approach and facilitate model development. This standardized approach enabled the consistent processing of data in all models and comparison of the predictive performance of all the models.

4.6 Evaluation Strategy

Some of the criteria used to evaluate the models will include the following:

- ROC-AUC Score: This will be the main assessment parameter to show whether the model has the ability to segregate the classes of low attention and normal attention.
- Precision, Recall and F1-Score: The measures will be applied to ensure the model accuracy in classification.
- Confusion Matrix: It is a method that will be applied in deciding what happens to the chosen model in prediction.

4.7 Selection Model

DoI: doi.org/10.63949/crinfo.v2i2.001

The final model will be the one that scores the highest on the test data on the model with highest ROC-AUC. In order to make interpretation of the results, feature importance analysis will be conducted on the models in order to establish factors influencing the attentional drop.

4.8 Implementation Environment

All of the methodology is executed in the Python language in a notebook environment. Libraries: There are libraries like Scikit-learn and XGBoost which are utilized in creating models, preprocessing and evaluation.

5. Results And Analysis

The below table illustrates the comparison between the performances of the models used:

Table 1. Performance Comparison of the models.

Model	ROC-AUC
Logistic Regression	**0.5835**
Random Forest	**0.5786**
XGBoost	**0.5795**

The model with the highest ROC-AUC was the Logistic Regression model with the highest value at **0.5835**.A subsequent evaluation of the classification measures showed that there was an overall accuracy of **55**%. Namely, the recall rates of the low-attention category in the model reached **0.59** which is far more impressive than the almost zero recall before class balancing was taken into consideration.
Technically, the fact that Logistic Regression was superior to the more complicated ensemble classifier, such as XGBoost.

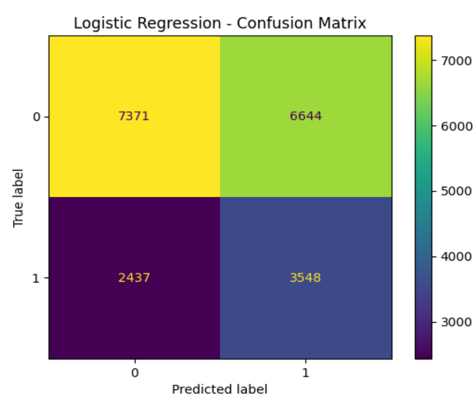

Figure 2.Confusion matrix of the Logistic Regression model showing classification performance, with true negatives (7,371), false positives (6,644), false negatives (2,437), and true positives (3,548).

5.1 Classification Report

DoI: doi.org/10.63949/crinfo.v2i2.001

Precision, recall, F1 score and overall accuracy were used to assess the Logistic Regression model. The performance measures are as summarized below:

Table 2. Performance Evaluation Metrics of the Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
0	0.75	0.53	0.62	14,015
1	0.35	0.59	0.44	5,985
Accuracy	–	–	0.55	20,000
Macro Average	0.55	0.56	0.53	20,000
Weighted Average	0.63	0.55	0.56	20,000

The model is able to achieve an overall accuracy of 55% on a sample set of **20,000** samples. There is moderate class imbalance as Class 0 has **14,015** samples and Class 1 has **5,985** samples.
In Class 0, the model reaches a high precision of **0.75**, which means that it provides high accuracy in classifying the instances as Class 0. Nevertheless, the recall value of **0.53** indicates that almost **50**% of the true Class 0 are falsely labeled.
In Class 1, the recall of **0.59** is comparatively better than the recall of Class 0 which is a sign of better detection of positive cases. However, the accuracy of **0.35** is significantly smaller, which indicates that the amount of false positives is large. The minority class had a lower F1 -score of **0.44** that indicates reduced overall performance.

5.2 Confusion Matrix Analysis

Confusion matrix gives more information about the classification behaviour of the model:

Table 3. Confusion Matrix of the Logistic Regression Model

	Predicted 0	Predicted 1
Actual 0	7,371	6,644
Actual 1	2,437	3,548

From the confusion matrix:

- True Negatives (TN) = **7,371**
- False Positives (FP) = **6,644**
- False Negatives (FN) = **2,437**
- True Positives (TP) = **3,548**

The model is right with the classification of **7,371** negative and **3,548** positive. Nonetheless, the number of false positive cases is quite high (6,644), which improves the accuracy of Class 1.

DoI: doi.org/10.63949/crinfo.v2i2.001

The middle-level recall of positives is presented by the relatively low number of false negatives (2,437). However, the confusion matrix shows that the model is prone to a large number of false positives, which shows a weak discriminative ability of the two classes.

5.3 Analytical Summary

Logistic Regression model shows the moderate predictive accuracy with observable discrepancy in effectiveness of classes. Although it has a good recall on the positive class, the high false-positive rate has a great influence on precision. Such results indicate that the model can best distinguish between the two classes even though the model can identify positive cases.

6. Insights of the Study

The tree-based machine learning models using feature importance analysis revealed some behavioural traits linked with attention lapses in the students. Of these, the video completion percentage stood out as one of the most significant, signifying how far students will go to interact with and complete educational material. Additionally, the usage time of the daily app was identified as a significant factor, which accounts for how much time learners spend using the app on a daily basis and the level of their participation. In addition, the frequency of sessions also had an impact on attention, as patterns of inconsistent or fewer sessions were likely to be associated with less learner attention. The results indicate that the behaviour of attention is significantly affected by the learners' dedication towards content consumption, frequent use of the platform, and continued involvement in learning activities. These findings can help educational software developers to pinpoint the learners who may be at risk and provide early interventions to help them succeed in their learning.

7. Limitations

There are numerous limitations to this study that need to be taken into account when assessing the results. One limitation is that a direct measure of learner attention was not utilized. Instead, an indirect measure using behavior-based interaction indicators was used; therefore, it is plausible that this may not provide an accurate reflection of all of the cognitive processes that contribute to learner attention. Another limitation is that much of the data was collected in a static manner and combined with other data to form an overall typology of behavioral characteristics. This static nature of the data restricts the model's ability to display longitudinal trends (temporal data) and patterns of learner engagement over time. In addition to the limitations of the data, the moderate ROC-AUC scores observed in this study indicate that the attention dynamics identified in digital learning environments may have only been partially captured by interaction indicators alone. Additional factors such as cognitive, emotional, and contextual factors are also important contributors to learner attention; therefore, these factors were not included in this study. The authors recommend that future research employs more specific behavioral indicators (temporal data) and utilizes

DoI: doi.org/10.63949/crinfo.v2i2.001

multiple data sources (multimodal data) in order to improve prediction accuracy related to learner attention loss.

8. Future Work

In order to improve the ability of future systems to predict the performance and usability of attention models in online learning contexts, future studies may focus on using Long Short-Term Memory (LSTM) network architectures. LSTM networks can be trained to recognize both temporal relationships among learners and changing patterns of learner behavior by using their ability to model temporal relationships over time, therefore enabling more effective identification of shifts in learner attentiveness. Additionally, the use of detailed clickstream data from online courses and other forms of remote learning would further provide insights into the ways in which learners engage with materials through active and passive learning styles. For example, real-time applications of these technologies (i.e., uniformed models and automatic delivery of intervention and re-engagement techniques via automated notifications) should be further examined as a way to increase learner engagement and to support learners when they fall below established thresholds of learner attention. Ultimately, these advances will create more engaging learning experiences, improved learning outcomes, and a more adaptable and intelligent digital learning environment.

9. Conclusion

This study introduced an interaction-based machine learning approach aimed at predicting student attention drop-out in online educational environments by analyzing non-intrusive behavioral data. To identify patterns linked to decreased student attention, a variety of machine learning models were evaluated, such as logistic regression, random forest, and XGBoost. The logistic regression model achieved the highest ROC-AUC score of 0.5835, indicating its moderate ability to differentiate between low-attention and normal-attention learners compared to the other models analyzed. Feature importance analysis revealed that video completion percentage, daily application usage time, and session frequency emerged as some of the most significant factors influencing learner attention. These findings emphasize the effectiveness of behavioral interaction data in monitoring student engagement and informing early intervention efforts. Future research could improve attention prediction systems by incorporating temporal learning models, conducting thorough clickstream data analyses, and developing adaptive intervention strategies in real-time, thus refining the accuracy and utility of these systems for digital learning.

DoI: doi.org/10.63949/crinfo.v2i2.001

ISSN: 3049-3412 (Online)
Volume No: 02 – Issue No: 02 – June 2026

References

[1] T. Anderson and F. Elloumi, "Theory and Practice of Online Learning," Athabasca University, 2004.
[2] G. Siemens, "Learning Analytics: The Emergence of a Discipline," American Behavioral Scientist, vol. 57, no. 10, pp. 1380–1400, 2013.
[3] R. S. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.
[4] D. B. West, "Introduction to Machine Learning for Predictive Data Analytics," MIT Press, 2014.
[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[7] J. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man, and Cybernetics, vol. 40, no. 6, pp. 601–618, 2010.
[8] Digital Learning Analytics Dataset, Kaggle, 2023. [Online]. Available: https://www.kaggle.com/

DoI: doi.org/10.63949/crinfo.v2i2.001